Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: A quantitative analysis

Lei Nie 1#, Gang Wu 2#, Weiwen Zhang 3*

1 Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown university, Washington DC 20057. 2 Department of Biological Sciences, University of Maryland at Baltimore County, Baltimore, MD 21250; 3 Microbiology Department, Pacific Northwest National Laboratory, P.O. Box 999, Mail Stop P7-50, Richland, WA 99352, USA.

# These two authors contributed equally to this paper.

* To whom correspondence should be addressed:

Weiwen Zhang, Ph.D.
Microbiology Department
Pacific Northwest National Laboratory
P.O. Box 999, Mail Stop P7-50, Richland, WA 99352
Phone: (509)-376-2016; Fax: (509)-372-1632; Email: Weiwen.Zhang@pnl.gov
Abstract

The modest correlation between mRNA expression and protein abundance in large scale datasets is explained in part by experimental challenges, such as technological limitations, and in part by fundamental biological factors in the transcription and translation processes. Among various factors affecting the mRNA-protein correlation, the roles of biological factors related to translation are poorly understood. In this study, using experimental mRNA expression and protein abundance data collected from Desulfovibrio vulgaris by DNA microarray and LC-MS/MS proteomic analysis, we quantitatively examined the effects of several translational-efficiency-related sequence features on mRNA-protein correlation. Three classes of sequence features were investigated according to different translational stages: i) initiation: Shine-Dalgarno sequences, start codon identity and start codon context; ii) elongation: codon usage and amino acid usage; and iii) termination: stop codon identity and stop codon context. Surprisingly, although it is widely accepted that translation initiation is the rate-limiting step for translation, our results showed that the mRNA-protein correlation was affected the most by the features at elongation stages, i.e., codon usage and amino acid composition (5.3-15.7% and 5.8-11.9% of the total variation of mRNA-protein correlation, respectively), followed by stop codon context and the Shine-Dalgarno sequence (3.7-5.1% and 1.9-3.8%, respectively). Taken together, all sequence features contributed to 15.2-26.2% of the total variation of mRNA-protein correlation. The study provides the first comprehensive quantitative analysis of the mRNA-protein correlation in bacterial D. vulgaris, and adds new insights into the relative importance of various sequence features in prokaryotic protein translation.
Keywords: Correlation; mRNA; Protein; Translational efficiency; Multiple regression

Running title: mRNA-protein correlation in *D. vulgaris*

Abbreviation: AA, Amino acid usage; CMR, Comprehensive Microbial Resource; CR, Condon usage; DVU, *D. vulgaris*; MFE, Minimal free energy; POD, Percent of differences; RBS, Ribosomal binding site; RP, Ribosomal protein; RSCU, Relative synonymous codon usage; SD, Shine-Dalgarno sequence; SRB, Sulfate-reducing bacteria;
Introduction

High throughput post-genomic technologies such as microarray and proteomics analyses have provided powerful methodologies to study patterns of gene expression and regulation at the genome scale (HORAK and SNYDER 2002; SMITH et al. 2002). Given the fact that no single approach can fully unravel the fundamental biology that is typically quite complex, an integrative approach of multiple levels of information is necessary and valuable to fully elucidate the complex biological system being studied. In several recent studies, integrative analyses of transcriptomic and translatomic data have helped researchers better understand the global regulatory processes and complex metabolic networks in living organisms (GYGI et al. 1999; GREENBAUM et al. 2002, GREENBAUM et al. 2003; HEGDE 2003; MOOTHA et al. 2003a; MOOTHA et al. 2003b; ALTER and GOLUB 2004).

Early large scale analyses of mRNA expression and protein abundance data showed that the correlation between mRNA and protein abundance was typically weak (FUTCHER et al. 1999; GYGI et al. 1999; IDEKER et al. 2001; GREENBAUM et al. 2003; WASHBURN et al. 2003). It has been proposed that three potential reasons for the lack of a strong correlation between mRNA and protein expression levels are: i) translational regulation, ii) differences in protein in vivo half-lives, and iii) the significant amount of experimental error, including differences with respect to the experimental conditions (GREENBAUM et al. 2003; BEYER et al. 2004). We recently performed a quantitative analysis of the contributions of various biochemical and physical sources to the correlation of mRNA and protein abundance in Desulfovibrio vulgaris. The results showed that analytic variations in mRNA expression and protein abundance contributed
to 34-44% of the total variation of mRNA-protein correlation, and protein and mRNA stabilities contributed to 5% and 2% of the total variation of mRNA-protein correlation, respectively (NIE et al. 2006a). However, since more than half of the variation remains unexplained, quantitative investigation of other factors, such as sequence features related to regulation at the translational level, will be important for further understanding of the mRNA-protein correlation.

Efficiency of protein biosynthesis depends on many factors. First, initial anchoring of ribosomes onto the mRNA depends on complementary binding of the Shine-Dalgarno (SD) sequence ~10 bases upstream of the start codon (SHINE and DALGARNO 1974), and a sequence close to the 3’ end of the 16S rRNA in the 30S ribosomal subunit (MCCARTHY and BRIMACOMBE 1994; STENSTROM et al. 2001). In this process, the short SD sequence motif serves as the ribosomal binding site (RBS). The ribosomal binding to this purine-rich SD region is of prime importance to locate the ribosome at the proper initiation codon (STENSTROM et al. 2001). The strength of this interaction has been used to estimate the efficiency of translation initiation (SCHURR et al. 1993; OSADA et al. 1999), or to predict the actual start sites (SUZEK et al. 2001).

Second, non-random use of synonymous codons in the coding region of highly expressed Escherichia coli genes indicates that sequences further downstream of the start codon could be of importance for translation efficiency (FAXEN et al. 1991; MCCARTHY and BRIMACOMBE 1994; COLLINS et al. 1995). A difference in translation rate in the order of 6-fold was found when infrequent and common codons were compared (SORENSEN et al. 1989). The 12 codons could impose a 15-fold difference in gene expression in studies using a lacZ based fusion system preceded by a canonical Shine-
Dalgarino sequence (LOOMAN et al. 1987; STENSTROM et al. 2001). Third, translation efficiency also depends on the availability of various amino acids. Among 20 amino acids, costs of synthesis vary from 12 to 74 high-energy phosphate bonds per molecule (AKASHI and GOJOBORI 2002). The selective advantage of using a cost-efficient amino acid in a highly expressed protein can be great. The evidence of natural selection of amino acid usage to enhance metabolic efficiency has been found in the proteomes of *E. coli* and *Bacillus subtilis* (AKASHI and GOJOBORI 2002). Fourth, translation termination depends upon the attachment of a release factor (RF) in the place of a tRNA in the ribosomal complex (ROCHA et al. 1999; KISSELEV et al. 2003). While release factors might exhibit different recognition efficiency to different stop codons, the pattern of stop codon usage changes more considerably than the start among prokaryotes (POOLE et al. 1995; ROCHA et al. 1999; OZAWA et al. 2002). In addition, the use of the stop codon was found to be correlated significantly with the G+C content of the genome, e.g., negatively for TAA and positively for TGA in *B. subtilis* (ROCHA et al. 1999). Moreover, cases have been reported in which the tri-nucleotides (stop codons) are not sufficient to terminate translation effectively (TATE et al. 1995, 1996). Studies showed that nucleotide distribution around the stop codons, especially the base following the stop codon, is significantly biased and is related to translation termination efficiency (BROWN et al. 1990; CAVENER and RAY 1991; TATE et al. 1996; POOLE et al. 1995, 1997).

*D. vulgaris* belongs to a group of obligate anaerobic microorganisms, sulfate-reducing bacteria (SRB, VOORDOUW 1996). Research interest in the SRB has been due to their corrosion of pipes and their ability to precipitate heavy metals (e.g. Cr$^{6+}$, Fe$^{3+}$,
U$^{6+}$) and radionuclides (e.g. U$^{6+}$) from solution via bacterial metal-reduction (HEIDELBERG et al. 2004). The genome of D. vulgaris was recently finished (HEIDELBERG et al. 2004). Our group has been using a whole-genome microarray and liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) to investigate gene expression and protein abundance in D. vulgaris under various growth conditions (ZHANG et al. 2006a, 2006b, 2006c, 2006d). Taking advantage of these experimental data, we attempted to determine the relative effects of various sequence features related to translational efficiency on mRNA-protein correlation in D. vulgaris in a single unified framework with a multiple regression approach. Sequence features considered in this study can be categorized into three classes with regard to the three stages of protein translation: 1) translation initiation: Shine-Dalgarno sequences, start codon identity and start codon context; 2) translation elongation: codon usage and amino acid usage; and 3) translation termination: stop codon identity and stop codon context. The results provided the first systematic quantitative analysis of the effects of translational efficiency-related sequence features on the mRNA-protein correlation and will help improve the understanding of mRNA-protein correlation, as well as regulation of gene expression at the translational level.

**METHODS**

**Datasets:** Three transcriptomic and translatomic datasets used for model construction and verification were collected from D. vulgaris DSM 644 grown on lactate- or formate-based chemically defined media. To minimize experimental variations between microarray and proteomic measurements, for each growth condition, identical cell
samples were used as starting materials to isolate RNA and proteins for analyses. The complete description of experimental design and cultivation conditions can be found in our previous study (ZHANG et al. 2006a). 1) **Microarray data:** Oligonucleotide microarray was used to obtain gene expression data (ZHANG et al. 2006a). To minimize variations between microarray measurements, for each growth condition, at least two replicate cell samples were used as starting materials to isolate RNA for analyses. In addition, each replicate sample used for RNA profiling was a pool of three individual biological replicates. Microarrays containing 3548 ORFs of *D. vulgaris* were designed by NimbleGen System Inc. (Madison, WI) using its Maskless Array Synthesizer (MAS) technology (NUWAYSIR et al. 2002; HEIDELBERG et al. 2004). Arrays were designed with JazzSuite software and the MAS units were used to manufacture the custom arrays. For each ORF, thirteen unique 24-mer oligonucleotides from throughout the ORF were printed onto glass microscope slides (NUWAYSIR et al. 2002). The complete description of data acquisition and analysis can be found in our previous study (ZHANG et al. 2006a). 2) **Proteomics data:** The *D. vulgaris* samples were analyzed by LC-MS/MS on a Finnigan model LTQ ion trap mass spectrometer (ThermoQuest Corp., San Jose, CA). MS analysis was performed using a Finnigan model LTQ ion trap (ThermoQuest Corp., San Jose, CA) with electrospray ionization (ESI) (ZHANG et al. 2006c). Peptide identification was performed using SEQUEST Version 2.7 (ThermoFinnigan, San Jose, CA) (ENG et al. 1994; YATES et al. 1995) to search the *D. vulgaris* protein sequence database (HEIDELBERG et al. 2004). The relative protein abundance was estimated based on the number of peptide hits (GAO et al 2003; QIAN et al 2005). The peptide hits for a given protein was the average of three LC-MS/MS measurements. The complete
description of proteomic data acquisition and analysis can be found in our previous study (ZHANG et al. 2006c, 2006d). In this study, to further remove the possible systematic errors caused by gene length and amino acid composition, we normalized the peptide hit by dividing it with a correction factor: “effective number of peptides”. The “effective number of peptides” is derived as follows: i) firstly we used the “PeptideCutter” program (http://ca.expasy.org/tools/peptidecutter/) to analyze every protein in the D. vulgaris genome and determined the possible cleavage sites by trypsin; ii) secondly we wrote a perl script to determine the number of peptides of all sizes after trypsin cleavage (the perl script available upon request); and iii) finally we added up the number of peptides of length 7 amino acids to 25 amino acids as the ‘effective number of peptides’ because the mass (m/z) detection range can only pick up peptides of this size range (ZHANG et al. 2006c, 2006d). The normalized peptide hits were then used as the protein abundance for all statistical analyses performed in this study.

In addition, to exclude the possibility that sequence features are surrogate measures of mRNA levels when there was a correlation between sequence features and measurement errors of mRNA, a regression model was trained to predict the mRNA abundance in one replicate with that in other replicates plus sequence features. We gradually lowered the threshold (fold-of-change) in the training sets of replicates until sequence features no longer contributed to the prediction. Thus the sequence features were completely independent of mRNA abundance of genes used in the following statistical analysis. The threshold values were determined as 3.5, 4.2 and 4.3 (fold change among replicates) for Lactate-Exponential (LE), Formate-Exponential (FE) and Lactate-Stationary (LS) conditions, respectively. Using this method, 349, 395, and 357
genes/proteins that satisfy the quality control criteria for LE, FE and LS condition, respectively, were used in the following multiple regression analyses.

**Quality of microarray and proteomic data and their correlation**: The quality of the microarray data was evaluated with the *Pearson* correlation coefficient analysis among multiple measurements. *Pearson* correlation coefficients of the microarray experiments are from 0.97 to 0.99 among replicate samples (NIE *et al*. 2006a), and *Pearson* correlation coefficients of LC-MS/MS measurements normalized by amino acid composition are greater than 0.86-0.92 among replicates, indicating good reproducibility. Correlation of mRNA expression and normalized protein abundance of cells grown was modest of about 0.54-0.63 (*p* value<0.001) by *Pearson* correlation coefficient for all conditions. The correlation levels were close to that previously reported for yeast (IDEKER *et al*. 2001).

**Cellular function category**: The cellular functional category of all genes in the *D. vulgaris* genome are downloaded from the Comprehensive Microbial Resource (CMR) of TIGR ([http://cmr.tigr.org](http://cmr.tigr.org)) (HEIDELBERG *et al*. 2004). Based on the original annotation, the genes/proteins are classified into 19 cellular functional categories.

**Identification and analysis of Shine-Dalgarno sequences**: Two different methods were used to identify the Shine-Dalgarno (SD) sequences in *D. vulgaris* genes. 1) *Free energy based method* (SCHURR *et al*. 1993; OSADA *et al*. 1999): this method aligned the 3’ end of 16S rRNA and 5’-UTR of an mRNA, then used a dynamic programming
algorithm to find the minimum free energy of a window of specific size (OSADA et al. 1999). Initially, we performed an analysis similar to what was described by OSADA et al. (1999), where the base-pairing potentials between the 3’ tail of 16S rRNA and 5’ UTR of all genes were calculated and averaged by positions to view the overall trend in the whole genome. Sequences “cctgcgctggatcctcttt” and “cctgcggttgatcctctta” from the 3’ end of 16S rRNA were used in D. vulgaris (NC_002937 and NC_005863) (HEIDELBERG et al. 2004) and E. coli (U00096) to calculate the free energy values in these two species, respectively. The C programs used to perform the calculation were kindly provided by Dr. Osada of the Institute for Advanced Biosciences of Keio University. To determine the effects of SD sequence during protein translation, we calculated the free energy for base-pairing of 16S rRNA with SD sequence for each gene. To do this, we extracted the 25-base and 50-base nucleotide sequence immediately upstream of the start codon of each gene. Each extracted sequence was aligned with the 3’-tail of 16S rRNA to compute the minimal free energy (MFE) with the Java Applet at http://www.mag.keio.ac.jp/~rsaito/Research/BasePAP/BasePAP.html (implemented by Dr. Osada. A modified JAVA code to batch-calculate MFE is available upon request). Since it is possible that various lengths of 16S rRNA tails used in the calculation might affect the accuracy of the MFE values, three lengths (13, 20 and 23) were used, which corresponded to sequences “gatcactccttt”, “gcgggtgatcctccttt” and “cctgcgctggatcctccttt”, respectively. 2) Probabilistic method (SUZEK et al. 2001): this method used a ‘seed’ sequence to train a probabilistic model of SD sequences, which was then used to find the SD sequences in regions upstream of start codons of all genes. A good seed sequence is the 3’-end of the 16S rRNA (SUZEK et al. 2001). Five copies
of 16S rRNA genes have an identical 3’-tail in *D. vulgaris* (Notably the annotated sequences were incomplete and missing the sequence “ctggatcacctccttt” at the 3’ end. However, this sequence does exist in all five copies of *D. vulgaris* 16S rRNA genes). Interestingly, although *D. vulgaris* is a GC-rich species, the sequence of the 16S rRNA 3’-tail is highly similar to most other prokaryotes (compared with other species described by MA *et al.* 2002) and it contains the typical anti-SD sequence “ctcct”, which complements with the default seed sequence “aggag” used in the RBSFinder program (downloaded at [ftp://ftp.tigr.org/pub/software/RBSfinder/](ftp://ftp.tigr.org/pub/software/RBSfinder/)). Two window sizes, 25 and 50, were used to search for SD sequences. The RBSFinder code was slightly modified to output the maximal RBS score (the modified perl code is available upon request).

**Start codon, stop codon and their contexts:** The identity of each start codon and stop codon was treated as a categorical variable during multiple regression analysis. The start codon context was defined as the upstream 30 bases and downstream 9 codons of start codon. Therefore, each sequence of start codon context is 60 bases long, including the start codon. To evaluate the potential of each start codon context to form a stable mRNA secondary structure, the minimum free energy of this region was computed with Vienna package *RNAfold* (HOFACKER 2003; HOFACKER and STADLER 2006). Stop codon and the base immediately downstream the stop codon was regarded as stop codon context. Each combination was treated as a categorical variable in multiple regression analysis described below.

**Analyses of the overall codon usage and amino acid usage:** The major trends in codon
usage and amino acid usage were revealed with a correspondence analysis. The relative synonymous codon usage (RSCU) was used in the correspondence analysis to remove the effects of amino acid usage. For amino acid usage, the raw codon counts were added up for each amino acid and used as input in the correspondence analysis. The CodonW software (http://codonw.sourceforge.net) was used for the correspondence analysis, generating four major axes accounting for most of the variations in codon usage or amino acid usage of D. vulgaris genes or proteins, respectively (WU et al. 2006b).

**Analyses of the preferences in codons and amino acids:** D. vulgaris is a GC-rich species (GC: 64%; GC3: 77%) (HEIDELBERG et al. 2004). Therefore, the unequal usage of amino acids or synonymous codons can be simply a result of mutational bias (KNIGHT et al. 2001). To examine the preferences for certain codons or amino acids, we used the ‘percent of differences’ (POD) to indicate the preferences (positive values) and avoidance (negative values), which takes the mutational bias into account. The POD is defined as:

\[
POD = 100 \times \frac{\text{Observed\_Count} - \text{Expected\_Count}}{\text{Expected\_Count}}
\]

where Observed\_Count is the actual number of each codon or amino acid in a set of genes, and Expected\_Count is the expected number of that codon or amino acid when codon usage and amino acid usage are solely results of base composition due to mutational bias. The base composition was estimated with the frequency of each base at the third codon position with the whole genome data because this position is the most neutral (SUEOKA et al. 1988). The *absolute* expected frequency of each codon is calculated as the product of the mutational bias of each base that composes the codon.
The *relative* expected frequency of each codon equals the *absolute* expected frequency of this codon divided by the sum of *absolute* expected frequencies of all codons in the same codon family, which contains all synonymous codons coding for the same amino acid. As a result, the relative expected frequency of codon ATG is 1 because there is only one codon coding for methionine. The same thing happens to codon TTG. Finally, the expected number of a codon equals the product of its *relative* expected frequency and the observed total number of codons of its codon family.

Preferences in amino acid usage were measured in a similar way. For a particular amino acid, the observed number is the sum of the number of codons coding for it. Its absolute expected frequency is the sum of the absolute expected frequency of its coding codons. Its relative expected frequency equals its absolute expected frequency divided by the sum of absolute expected frequencies of all amino acids (essentially 61 codons, because three termination codons were excluded). Finally, the expected count of this amino acid is the product of its relative expected frequency and the total number of amino acids observed in a set of genes.

The significance of the POD was evaluated with the following method: the observed count for a particular codon was assumed to follow a binomial distribution $B(n,q)$, in which $n$ is the total number of codons in its codon family, and $q$ is the relative expected frequency of the codon within its synonymous codon family. The significance is based on testing whether the observed relative frequency equals the expected relative frequency of this codon in its synonymous codon family. Since the number $n$ is a large number here, a normal approximation has been used to approximate the binomial distribution (OTT and LONGNECKER 2001; DEVORE and FARNUM...
Because multiple tests were involved, a *Bonferroni* procedure has been implemented to adjust the *p*-values obtained from multiple tests (OTT and LONGNECKER 2001). For an amino acid, *p*-values were computed in a similar way to test whether an amino acid count follows the distribution $B(n, q)$, where $n$ is the total number of amino acids and $q$ equals the relative expected frequency of this amino acid.

**Correlation and regression:** Correlation coefficients, such as *Pearson’s* correlation coefficient and Spearman rank correlation coefficient, were computed (DEVORE and FARNUM 2005). Furthermore, single regression and multiple regression analyses were performed to measure correlation pattern between mRNA and protein abundance as described before (OTT and LONGNECKER 2001; NIE et al. 2006a). To obtain a reliable correlation between mRNA and protein abundance, only proteins with variations among measurement replicates less than three fold were included. Fold change in original scale is equivalent to the arithmetic difference in the log scale, also called range of samples (MONTEGOMERY 2001). For instance, $\max(y_1, y_2, y_3)/\min(y_1, y_2, y_3)=3$ is equivalent to $\log[\max(y_1, y_2, y_3)]-\log[\min(y_1, y_2, y_3)]=\log(3)$. Previously, we reported the proteomic-mRNA correlation through $R^2_{\text{mRNA}}$ from a simple regression:

$$y_i = \alpha + \text{mRNA}_i \times \beta$$

where $\text{mRNA}_i$ is the log of mRNA expression level for the $i$th gene (NIE et al. 2006a). In this study, we included the quantitatively measured sequence features into a multiple regression:

$$y_i = \alpha + \text{mRNA}_i \times \beta + \sum_{j=1}^{k} \beta_j x_{ij},$$

where $x_{ij}$ is the quantitatively measured sequence feature for the $i$th gene and the $j$th feature.
where $x_{ij}$ refers to $j$th covariate (measuring sequence feature such as codon usage, $k$ is the number of covariates of a particular sequence feature) of the $i$th gene, $\beta_j$ represents the slope for the $j$th covariate (NIE et al. 2006a). Particularly, we reported the

$$\frac{R_{mRNA,sequences}^2 - R_{mRNA}^2}{1 - R_{mRNA}^2}$$

as the adjusted $R^2$ for the mRNA-protein correlation. For each covariate, we used the standard F-test to examine whether they were significant ($p$-value of the F-test reported) (OTT and LONGNECKER 2001). Finally, since the variation from each of these effects may not be additive, we investigated their joint effects on the mRNA-protein correlation in a single multiple regression analysis by using the following equation:

$$y_i = \alpha + \text{mRNA}_i \times \beta + \sum_{j=1}^{m} \beta_j x_{ij}$$

(4)

where all sequence features were included as covariates ($m$ is the total number of all covariates). $P$-values associated with each covariate were measured.

RESULTS

1. Sequence features related to translation initiation

1.1 Shine-Dalgarno sequences

To determine the effects of SD sequences, we need to be certain that the same mechanism of translation initiation also exists in $D. vulgaris$. To confirm this, we systematically calculated the average free energy of the potential base pairing of 16S rRNA 3’-tail and upstream of $D. vulgaris$ genes with the free-energy-based method (OSADA et al. 1999). A sharp drop of the free energy was observed in the region of -25 ~ -5 in all genes in the
*D. vulgaris* chromosome and mega-plasmid (Figure 1). This trend is highly similar to what has been observed in *E. coli* (OSADA *et al.* 1999), suggesting Osada’s method works very well in *D. vulgaris* as well. It also indicates that the same mechanism of translation initiation is employed in *D. vulgaris* as in most prokaryotes. Interestingly, the average free energy by position is lower in *D. vulgaris* than in *E. coli* (Figure 1), which could be due to the high GC composition of the *D. vulgaris* genome.

Using the same free-energy-based method (OSADA *et al.* 1999), we calculated the minimum free energy (MFE) of 16S rRNA-SD base pairing for all *D. vulgaris* genes. As one might expect, the length of the 16S sRNA 3’-tail and upstream sequences might affect the free energy value calculated. To test this effect, we have chosen three lengths (13, 20 and 23) of 16S rRNA 3’-tail and two lengths (25 and 50) of mRNA upstream sequences, and calculated the MFE for various combinations. In addition to the free-energy-based method, we have also employed a probabilistic method to quantify the strength of RBS (SUZEK *et al.* 2001). The putative SD sequence was found with a RBS score higher than the threshold value (SUZEK *et al.* 2001). The result showed that the contribution varied considerably among various methods of SD MFE calculations, with MFE20_50 (20 refers to the length of 16S rRNA 3’-tail and 50 refers to the length of the upstream sequence of genes) explaining the highest variations in mRNA-protein correlation (up to 1.9-3.8% among various datasets, p-value <0.001; Table 1, Supplementary Table 1). The result also showed that the SD effect calculated by probabilistic method does not seem to explain much of the variations in mRNA-protein correlation (mostly around 1.0%, Supplementary Table 1). In addition, only a weak correlation between RBS score and MFE20_50 was observed (r = -0.27, p<0.0001),
suggesting that these two methods may be different in evaluating SD sequences. In the
study hereafter, MFE20_50 was used in all analyses involved.

A recent study in *E. coli* showed that the SD MFE values in highly expressed
genes (defined as genes whose protein product can be detected with on 2D gels) were
lower than in other genes (LITHWICK and MARGLIT 2003). We have previously found
that the *D. vulgaris* proteins detected by LC-MS/MS represented the highly expressed
genes (Supplementary Figure 1; ZHANG *et al.* 2006c). When the MFE20_50 value of the
detected proteins was compared with that of all proteins in the *D. vulgaris* genome, a
similar frequency distribution pattern was found (Figure 2), suggesting that overall there
was no strong evidence for lower MFE in proteins identified in LC-MS/MS.

We have previously found that mRNA-protein correlation may be different
among various functional categories (NIE *et al.* 2006a). Given the fact that MFE values
may be related to mRNA expression and protein abundance, one immediate question is
whether MFE values also varied among functional categories of a given gene/protein.
The results showed that while most of the functional categories shared similar levels of
MFE values, genes from several functional categories, such as “amino acid biosynthesis”,
“central intermediary metabolism”, “energy metabolism”, “protein fate” and “protein
synthesis”, had lower MFE values (Figure 3-I). This pattern appears to be consistent
when we used corresponding genes of the proteins identified under the three growth
conditions examined in this study (Figure 3-II, 3-III and 3-IV).

1.2. Start codons

In the *D. vulgaris* genome, the start codon ATG is the most frequent start codon,
consistent with the early conclusion that ATG is a preferred start codon, independent of the G+C content (ROCHA et al. 1999). Approximately ~82% of the *D. vulgaris* genes start with this codon, while the less frequently used TTG and GTG codons are found in 5.4% and 13% of the genes (Figure 4A). This strong bias in start codons was also observed in the proteins detected in various growth conditions (Figure 4A). This observation confirms that the canonical start codon ATG is more translationally optimal than non-canonical start codons. However, overall, the start codon identity explained only 0.1-0.7% of the total variation in mRNA-protein relationship under the three growth conditions as indicated by regression analyses (Table 1).

1.3. *Start codon context*

Studies showed that stem-loop structures formed at the start site can affect the accessibility of the SD sequence or start codon for ribosomal binding (de Smit and VAN DUIN 1990, 1994; ROCHA et al. 1999). To determine the potential effects of these mRNA secondary structures on protein translation, we computed the minimum free energy for the 60-base sequences spanning the start codon with *RNAfold* (HOFACKER 2003; HOFACKER and STADLER 2006). We found that proteins detected in our study tend to have relative high MFE values compared with all proteins in the *D. vulgaris* genome (Figure 4B); suggesting that avoidance of mRNA secondary structure might be a strategy for genes to achieve a high translation rate. Overall, the start codon context explains 0.3-2.5% of the variation in mRNA-protein correlation under the three growth conditions (Table 1), which is larger than that by start codons alone.
2. Sequence features related to translation elongation

2.1. Codon usage pattern

The codon usage in the G+C rich *D. vulgaris* genome has not been fully investigated before (HEIDELBERG *et al*. 2004). In this study, two approaches have been employed to investigate the unequal codon usage in *D. vulgaris*. The first approach is to compare the extent of codon usage deviated from the expected frequency between detected proteins and all proteins to determine which codon is associated with protein translation. Due to the high GC composition of the *D. vulgaris* genome, the observed unequal frequency of synonymous codons can be simply a result of biased base composition due to mutational bias (KNIGHT *et al*. 2001). Therefore, we used “percent of differences” (POD) to measure how much the observed codon frequency deviated from the expected frequency determined based on the base composition alone. Briefly, a positive POD value suggests an over-representation of the codon, whereas a negative POD indicates an under-representation. If the codon usage is associated with gene expression level, we would expect to see significantly different POD values between detected proteins and all proteins. Indeed, we found that except for Glu and Val codon families, the POD values of at least one codon within all other 16 sense codon families (Met and Trp are excluded because they only have one codon) were significantly different (*p* values <0.0001) between detected proteins and all proteins (Figure 5A and 5B). In most cases, the absolute values of POD are greater in detected proteins than in all proteins, suggesting a stronger bias for these codons in detected proteins. This observation is consistent with previous findings in *E. coli* and *S. cerevisiae* (IKEMURA 1981, 1982, 1985). Taken together, it is obvious that codon usage in *D. vulgaris* is strongly associated with protein
expression.

The second approach is a correspondence analysis of the relative synonymous codon usage (RSCU). The first four major trends in codon usage (CR1-4) determined by correspondence analysis accounted for 17.1%, 4.4%, 3.9% and 3.7% of the total variations in RSCU of all *D. vulgaris* genes (WU *et al.* 2006b). We included the first four major axes in codon usage of the *D. vulgaris* genome in a multiple regression analysis. The result showed that codon usage alone contributed to 5.3-15.7% of the total variation in mRNA-protein correlation under the three conditions (*p*-value <0.001) (Table 1).

2.2. Amino acid usage

It was observed that the usage of some amino acids was correlated with gene expression (AKASHI 2003; CHANDA *et al.* 2005; SCHABER *et al.* 2005; AKASHI 2003). To determine whether it is same case for amino acid usage in *D. vulgaris*, we applied a similar approach to the one used for codon usage, as described in the previous section. The only difference was that this time the frequencies of all synonymous codons were summed. Therefore, if the usage of an amino acid is associated with protein expression, we expect the POD value of this amino acid in detected proteins to be different from that in all proteins. As expected, the POD values of amino acids Asn, Asp, Glu, Ile, Tyr and Lys in detected proteins are significantly higher than those in all proteins (*p*-values < 0.0001 for all these amino acids) (Figure 6), suggesting that these amino acids are preferred in more abundant proteins.

Given the fact that amino acid usage in *D. vulgaris* does associate with protein expression, we employed a multiple regression analysis to determine its relative
contribution to mRNA-protein correlation. A correspondence analysis of the deduced protein sequences of 3522 genes in *D. vulgaris* was performed to reveal the major trends. The first four trends (AA1-4) identified by correspondence analysis explain 18.6%, 11.5%, 9.7% and 7.0% of the variability in *D. vulgaris* amino acid usage, respectively (WU et al. 2006b). When we included the first four major axes in amino acid usage of the *D. vulgaris* genome in a multiple regression analysis, the result showed that amino acid usage alone contributed to 5.8-11.9% of the total variation in mRNA-protein correlation under the three growth conditions (*p*-value <0.001) (Table 1).

3. Sequence features related to translation termination

3.1. Stop codons

In *D. vulgaris*, three types of stop codons are all commonly used, with approximately 45%, 41% and 14% among all genes for TGA, TAG and TAA, respectively (Figure 7A). However, among the proteins detected under our three growth conditions, TAG seems to be the most preferred stop codon: about 56-59% of protein encoding genes use TAG as a stop codon (Figure 7A). Regression analysis showed that stop codon identity was an important feature affecting mRNA-protein correlation. It alone explained 1.3-2.3% of total variation in mRNA-protein correlation under the three growth conditions (Table 1).

3.2. Stop codon context

The base immediately downstream of the stop codon was also investigated for its role in translation termination. In *D. vulgaris*, “C” appears to be preferred after all three stop codons (Figure 7B). The most frequent stop signal is TAGC, which is even higher in
proteins identified under the three conditions, suggesting that it might be the optimal stop signal. Simple regression analysis confirmed that stop codon context was an important feature affecting mRNA-protein correlation. It alone contributed 3.7-5.1% of total variation in mRNA-protein correlation under the three growth conditions (Table 1).

4. Multiple regression analysis

In order to quantitatively determine the relative contribution of each translation efficiency-related feature on mRNA-protein correlation, all sequence features studied above were integrated into a multiple regression analysis. The results showed that these features together accounted for about 15.2-26.2% of the total variation at mRNA-protein correlation (Table 1). The $p$-values of this regression model, for all three conditions, are all less than 0.00001, which indicates contribution of these features to the mRNA-protein variation is significant. In addition, the result is very consistent in all three growth conditions.

To further evaluate the multiple regression model itself, and to verify the sincerity of the contribution resulting from this multiple regression, we ran two bootstrap tests by keeping sequence features unchanged for all genes, while randomly permutating their proteomic abundance among the genes so that the proteomic abundance of a given gene is randomly assigned to a different gene. The bootstrap tests were run by randomly selecting 1000 permutations for each test. For each permutation, a multiple regression was fitted and $R^2$ was reported as we did for the real data. The bootstrap $p$-value is reported as the probability that the simulated $R^2$ is larger than the $R^2$ associated with the real data. A smaller $p$-value suggests the $R^2$ obtained for the real model is statistically
more significant. The two null models for the bootstrap tests were that 1) the contribution by mRNA levels and all sequence features are not larger than mRNA level alone, and 2) the contribution by mRNA levels and all sequence features is no larger than that by mRNA levels and initiation-related sequence features (excluding elongation and termination related sequence features), respectively. The results showed the $p$-values from the first bootstrap analysis are 0 for the contributions computed in all three growth conditions, and the $p$-value for second null hypothesis is less than 0.0001 for all three conditions. The results demonstrated that correlation of mRNA expression and protein abundance was affected at a fairly significant level by multiple sequence features related to translational efficiency in *D. vulgaris*. Among them, the amino acid usage and codon usage are the top two factors, followed by stop codon context and SD sequences (Table 1).

**DISCUSSION**

It is widely accepted that gene regulation in prokaryotes occurs mainly at the transcription level and secondarily at the level of translation (LANGE and HENGGE-ARONIS 1994; CHHABRA *et al.* 2006). Nevertheless, so far no systematic quantitative analysis has been performed on the effects of various translation related sequence features on mRNA-protein correlation, and it remains unclear how strong the translation control is. In this study, we exploited the whole-genome mRNA expression and LC-MS/MS proteome abundance data from *D. vulgaris* grown in three conditions to gain insights into how the mRNA-protein correlation may be affected by various sequence features related to translation efficiency (ZHANG *et al.* 2006a; ZHANG *et al.* 2006c).
The major sequence features in each translational stage were quantified and their effects on mRNA-protein correlation were investigated. Multiple regression analyses of all sequence features showed that they together contributed up to 15.2-26.2% of the total variation of mRNA-protein correlation, suggesting that regulation at the translational level is indeed involved in determining mRNA-protein correlation in *D. vulgaris*. The result provides a valuable estimate of the contribution of translation related sequence features to protein synthesis. It may be worth noting that although several previous studies have been performed to determine the contribution of some of these sequence features to mRNA or protein abundance (POOLE *et al*. 1995; ROCHA *et al*. 1999; LITHWICK and MARGLIT 2003), these sequence features were not examined in a unified framework. To our knowledge, our analysis represents the first attempt to address the relative contribution of these sequence features to the mRNA-protein correlation with a single statistical model.

This investigation also benefited from the expression datasets we used. First, mRNA expression and protein abundance were determined for a pair of samples prepared in parallel to minimize the variations across experiments; second, the dataset is relatively large (more than 400 proteins and their corresponding genes); and third, three individual datasets were collected independently for three growth conditions (ZHANG *et al*. 2006a, 2006c; 2006d). All these have substantially contributed to the reliability of our analyses; therefore, the results from our analyses may represent a general phenomenon in *D. vulgaris*.

The traditional view on the contribution of SD sequences on mRNA-protein correlation was that the competition between the ribosome-RBS interaction and the
mRNA structure results in nearly “all or none” expression, leaving almost no room for regulation at the translation level (DE SMIT and VAN DUIN 1994; ROCHA et al. 1999). Therefore, all moderately or highly expressed genes should have a good combination of a good RBS and an efficient start codon. In contrast, some other findings suggested that the SD sequence has little influence on protein expression, as the binding of the 16S ribosomal subunit has been found not to be essential for translation initiation (CALOGERO et al. 1988). In fact, there exist mRNAs without SD sequences (BONI et al. 2001). In a recent study with E. coli and Haemophilus influenzae genomes, it was also proposed that the base-pairing potential of the mRNA SD sequence and the rRNA seemed to have a negligible effect on protein expression (LITHWICK and MARGLIT 2003). However, our study with D. vulgaris indicates that differential MFE values for SD sequences were present among gene/protein of different abundance level. Moreover, regression analysis showed that SD sequences contributed to 1.9-3.8% of the total variation of mRNA-protein correlation, suggesting that SD sequences do play a role in translational control.

It has been suggested that translation initiation is a rate-limiting step when compared with the elongation and termination stages of protein biosynthesis (Vellanoweth and Rabinowitz 1992, DE SMIT and VAN DUIN 1994, ROCHA et al. 1999; ROMBY and SPRINGER 2003). However, our study showed that features related to translation initiation (start codon and start codon context) play only a minor role in determining mRNA-protein correlation in D. vulgaris. To the contrary, the sequence features involved in translation elongation, such as codon usage and amino acid usage may be more important in determining mRNA-protein correlation in D. vulgaris. This
result is consistent with the conclusion by LITHWICK and MARGLIT (2003), who found codon bias had the greatest influence on protein expression levels, and also consistent with a recent study which estimated that codon bias accelerates translation in *E. coli* by no more than 60% in comparison to microbes with very little codon bias (DETHLEFSEN and SCHMIDT 2005).

Translation elongation can be retarded or stalled by the use of rare codons and the availability of various amino acids as building blocks (VARENNE et al. 1989, ROCHA et al. 1999, AKASHI and GOJOBORI 2002). The second major sequence feature that influences the mRNA-protein correlation was identified as the amino acid usage of the proteins in *D. vulgaris*. Metabolic cost of amino acids has been previously proposed to explain the preferences in amino acid usage (AKASHI and GOJOBORI 2002); abundant proteins tend to use more cost-efficient amino acids. However, no obvious correlation was found between amino acid preferences and its metabolic cost in our study. For instance, amino acids (Asn, Asp, Glu, Ile, Tyr and Lys) preferred in detected proteins are not of low metabolic cost (AKASHI and GOJOBORI 2002). Instead, we observed a strong correlation between protein hydrophobicity and AA2 (*Pearson* correlation coefficient 0.78), suggesting that selection for protein structure remains as the major determinant of amino acid usage in *D. vulgaris*.

It has been shown that the identity of the stop codon and stop codon context influence the rate of translation termination (POOLE et al. 1995). Several recent studies showed that the average gene expression level among the three stop codon groups in the five model organisms and humans does not differ significantly (*p* >0.05) (SUN et al. 2005), and the association of stop codons with high expression as reflected by the
coefficient was found to be weak (LITHWICK and MARGLIT 2003). Our analysis indicated that, while the contribution by stop codons was not dramatic (1.3-2.3% to the total variation of mRNA-protein correlation), the contribution by stop codon context was found as the third greatest sequence feature affecting mRNA-protein correlation, contributing 3.7-5.1% of the total variation. Therefore, the stop codon context may be more important in affecting the translation efficiency than previously suggested.

It might be noteworthy that while our previous study addressed the effects of experimental challenges and various physical properties of mRNA or proteins, such as analytic variation, stability of mRNAs and proteins, and the cellular functional category of genes/proteins, etc. (NIE et al. 2006a), this paper focuses on the impact of the sequences to the proteomic and mRNA correlation pattern. An attempt has also been made to integrate all the features tested in this work and our previous publication into one multiple regression model. Consequently we found that more than 71% of mRNA-protein correlation variation can be accountable (NIE et al. unpublished data), suggesting that we have identified most of the factors which can potentially affect mRNA-protein correlation. In addition, through F-test the results showed that the contributions by translation level factors are significantly independent of the contributions by errors and protein stability with $p$-value less than 0.0001 (OTT and LONGNECKER 2001; NIE et al. unpublished data). Although the model could be further improved if other related features, such as mRNA and protein degradation rate, can be included or larger proteomic datasets and their corresponding gene expression data become available, the results obtained here have greatly enhanced our understanding of correlation between mRNA expression and protein abundance. The results also lay down an important foundation for developing statistical
tools in integrating microarray and proteomic data (NIE et al. 2006b). In addition, the study extended the concept of integrated genomics by including experimental genomics data and other quantitative properties that are sequence-dependent. Without much modification, the methodology described in this paper can also be applied to the analysis of other model organisms such as *E. coli* and *S. cerevisiae*.

As a complicated biological process, many details of translation still need further investigation, for example, highly expressed protein may not necessarily require high abundance of mRNA if they have a translation rate higher than average. In this study, we made the initial attempt to use post-genomic data to analyze translation process in a simplified framework; it will be more informative to include other experimental parameters, such as RNA decay and protein degradation into the model once they become available in the future. We would like to point out that the conclusions reached were also subjected to limitation of datasets (*e.g.* quality of data and size of dataset). For example, the lack of significant contribution of start codons to variations in the mRNA-protein correlation may be due to the fact that most proteins used in this analysis already have a good start codon. In addition, the study was also subjected to limitation of our methods to quantify sequence features, such as minimal free energy of SD computation. Because of all these reasons, caution needs to be taken in further interpreting these conclusions as general rules in all prokaryotic cells. Nevertheless, our study provides the first comprehensive quantitative analysis of various sources contribution to the variation in the mRNA-protein correlation and shed some lights on the translation process in bacterial *D. vulgaris*. 
ACKNOWLEDGMENTS

We would like to thank the Environmental Molecular Sciences Laboratory (EMSL) at the Pacific Northwest National Laboratory (PNNL) for use of the proteomic instrumentation in generating data used in this research (EMSL proposal 15891). The EMSL is a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research. We are grateful to three anonymous referees for many constructive suggestions and comments, which greatly improved the quality of the paper. We are grateful for the valuable comments from Drs. Stephen J. Freeland, Philip Farabaugh, Janice Zengel, David Carlini and Marie desJardins. We are also grateful to Mr. Bryan Mackay for a critical reading of the manuscript. The research described in this paper was conducted under the Laboratory Directed Research and Development Program (to WZ) at the Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RLO1830. G. Wu was supported by the award 0317349 (To SJF) from the NSF DBI program Biological Databases and Informatics.
LITERATURES CITED


Faxen, M., J. Plumbridge and L.A. Isaksson. 1991. Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471-


Lange, R., and R. Hengge-Aronis. 1994. The cellular concentration of the σ^S subunit of RNA polymerase in *Escherichia coli* is controlled at the levels of transcription, translation, and protein stability. Genes Dev. 8:1600-1612.


Poole, E.S., R. Brimacombe and W.P. Tate. 1997. Decoding the translational termination
signal: the polypeptide chain release factor in *Escherichia coli* crosslinks to the base following the stop codon. RNA **3**: 974-982.


Tate, W.P., E.S. Poole, J.A. Horsfield, S.A. Mannering, C.M. Brown, et al. 1995. Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. Biochem Cell Biol. 73: 1095-1103.


Antonie van Leeuwenhoek. 89: 221-237.


Figure Legends:

**Figure 1:** Sharp decreases in free energy around position -25 and -5 (relative to start codon) are observed in the *D. vulgaris* (DVU) and the meta-plasmid of *D. vulgaris* (DVUA) of *D. vulgaris*, and *E. coli* (ECOLI). The free energy of all genes on each chromosome was averaged by position.

**Figure 2:** Frequency distribution of SD-rRNA interaction minimum free energy for all *D. vulgaris* proteins and proteins identified in three growth conditions. LL: lactate-log phase; FL: formate-log phase; LS: lactate-stationary phase.

**Figure 3:** Minimum free energy of SD-rRNA interaction of genes belonging to various functional categories. I) all genes; II) Lactate-log condition; III) Formate-Log condition; and IV) Lactate-Stationary condition. The cellular functional categories are: A: Amino acid biosynthesis; B: Biosynthesis of cofactors and carriers; C: Cell envelope; D: Cellular processes; E: Central intermediary metabolism; F: DNA metabolism; G: Disrupted reading frame; H: Energy metabolism; I: Fatty acid and phospholipid metabolism; J: Hypothetical proteins; K: Other categories; L: Protein fate; M: Protein synthesis; N: Biosynthesis of purines and pyrimidines; O: Regulatory functions; P: Signal transduction; Q: Transcription; R: Transport and binding proteins; S: Unknown function. The functional categories without any proteins detected are left blank in Figure 3-II, III and IV. The central, upper and low horizontal lines in the plots represent mean, plus and minus the standard deviation for each function category.
**Figure 4:** A) Frequency distribution of start codon usage in all *D. vulgaris* proteins and proteins identified from three growth conditions; and B) Frequency distribution of minimum free energy of predicted mRNA secondary structure at start codon context of all *D. vulgaris* proteins and proteins identified from three growth conditions.

**Figure 5:** Comparison of the preferences of synonymous codon usage between all *D. vulgaris* proteins and proteins identified from three growth conditions. A) Two-fold codon families; and B) Multi-fold codon families.

**Figure 6:** Comparison of the preferences of amino acid usage between all *D. vulgaris* proteins and proteins identified from three growth conditions.

**Figure 7:** A) Frequency distribution of stop codon usage in all *D. vulgaris* proteins and proteins identified from three growth conditions; and B) Frequency distribution of tetranucleotide termination signals in all *D. vulgaris* proteins and proteins identified from three growth conditions.

**Supplementary Figure 1:** Frequency distribution of mRNA abundance of all *D. vulgaris* proteins and proteins identified during lactate-log phase.
Table 1. Contribution of various features to the total variation of mRNA-protein correlation *.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Partial R Square (Lactate-Log)</th>
<th>Partial R Square (Formate-Log)</th>
<th>Partial R Square (Lactate-Stationary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA expression</td>
<td><strong>0.384</strong> ($p&lt;0.0001$)</td>
<td><strong>0.278</strong> ($p&lt;0.0001$)</td>
<td><strong>0.394</strong> ($p&lt;0.0001$)</td>
</tr>
<tr>
<td>SD sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFE20_50</td>
<td><strong>0.038</strong> ($p=0.0003$)</td>
<td><strong>0.031</strong> ($p=0.0005$)</td>
<td><strong>0.019</strong> ($p=0.0095$)</td>
</tr>
<tr>
<td>Start codon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATG</td>
<td>0.002</td>
<td>0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>TTG</td>
<td>0.002</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>GTG</td>
<td>0.002</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Sum of start codon **</td>
<td><strong>0.006</strong> ($p=0.5660$)</td>
<td><strong>0.001</strong> ($p=0.9175$)</td>
<td><strong>0.007</strong> ($p=0.4733$)</td>
</tr>
<tr>
<td>Start codon context</td>
<td><strong>0.018</strong> ($p=0.0131$)</td>
<td><strong>0.026</strong> ($p=0.0014$)</td>
<td><strong>0.003</strong> ($p=0.3016$)</td>
</tr>
<tr>
<td>Codon usage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR1</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>CR2</td>
<td>0.003</td>
<td>0.029</td>
<td>0.011</td>
</tr>
<tr>
<td>CR3</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>CR4</td>
<td>0.049</td>
<td>0.123</td>
<td>0.056</td>
</tr>
<tr>
<td>Sum of codon usage **</td>
<td><strong>0.053</strong> ($p=0.0011$)</td>
<td><strong>0.157</strong> ($p=0.0000$)</td>
<td><strong>0.067</strong> ($p=0.0001$)</td>
</tr>
<tr>
<td>Amino Acid usage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA1</td>
<td>0.030</td>
<td>0.021</td>
<td>0.011</td>
</tr>
<tr>
<td>AA2</td>
<td>0.017</td>
<td>0.000</td>
<td>0.017</td>
</tr>
<tr>
<td>AA3</td>
<td>0.023</td>
<td>0.063</td>
<td>0.026</td>
</tr>
<tr>
<td>AA4</td>
<td>0.003</td>
<td>0.035</td>
<td>0.005</td>
</tr>
<tr>
<td>Sum of amino acid usage **</td>
<td><strong>0.073</strong> ($p=0.0000$)</td>
<td><strong>0.119</strong> ($p=0.0000$)</td>
<td><strong>0.058</strong> ($p=0.0003$)</td>
</tr>
<tr>
<td>Stop codon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGA</td>
<td>0.020</td>
<td>0.020</td>
<td>0.008</td>
</tr>
<tr>
<td>TAG</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>TAA</td>
<td>0.003</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>Sum of stop codon **</td>
<td><strong>0.023</strong> ($p=0.0422$)</td>
<td><strong>0.023</strong> ($p=0.0292$)</td>
<td><strong>0.013</strong> ($p=0.2095$)</td>
</tr>
<tr>
<td>Stop codon context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAAA</td>
<td>0.004</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>TAAT</td>
<td>0.000</td>
<td>0.005</td>
<td>0.001</td>
</tr>
<tr>
<td>TAAC</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>TAAG</td>
<td>0.002</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>TAGA</td>
<td>0.000</td>
<td>0.000</td>
<td>0.009</td>
</tr>
<tr>
<td>TAGT</td>
<td>0.000</td>
<td>0.008</td>
<td>0.000</td>
</tr>
<tr>
<td>TAGC</td>
<td>0.001</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>TAGG</td>
<td>0.001</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>TGAA</td>
<td>0.005</td>
<td>0.001</td>
<td>0.012</td>
</tr>
<tr>
<td>TGAT</td>
<td>0.001</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>TGAC</td>
<td>0.016</td>
<td>0.018</td>
<td>0.017</td>
</tr>
<tr>
<td>TGAG</td>
<td>0.007</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Sum of stop codon context **</td>
<td><strong>0.037</strong> ($p=0.3669$)</td>
<td><strong>0.041</strong> ($p=0.1916$)</td>
<td><strong>0.051</strong> ($p=0.1079$)</td>
</tr>
<tr>
<td>Sum of all features above***</td>
<td>0.249</td>
<td>0.398</td>
<td>0.218</td>
</tr>
<tr>
<td>Total ****</td>
<td><strong>0.169</strong> ($p&lt;0.00001$)</td>
<td><strong>0.262</strong> ($p&lt;0.00001$)</td>
<td><strong>0.152</strong> ($p&lt;0.00001$)</td>
</tr>
</tbody>
</table>

* All scores are partial scores after mRNA, i.e., explanation of protein variations after taking away mRNA effects. The “Sums” listed for each group are sum of “partial R-squares” which has removed the overlapping effect among various variables.

** P-value measurements are from the standard F-test.

*** Not including mRNA expression.

**** The results are from multiple regression analyses that include all features.
**Supplementary Table 1.** Comparison of contributions of Shine-Dalgarno sequences by different methods.

<table>
<thead>
<tr>
<th>Shine-Dalgarno Sequences</th>
<th>Partial R Square (Lactate-Log)</th>
<th>Partial R Square (Formate-Log)</th>
<th>Partial R Square (Lactate-Stationary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFE13_25</td>
<td>0.024 (p=0.0038)</td>
<td>0.028 (p=0.0009)</td>
<td>0.013 (p=0.0298)</td>
</tr>
<tr>
<td>MFE13_50</td>
<td>0.020 (p=0.0078)</td>
<td>0.014 (p=0.0180)</td>
<td>0.006 (p=0.1387)</td>
</tr>
<tr>
<td>MFE20_25</td>
<td>0.028 (p=0.0018)</td>
<td>0.013 (p=0.0239)</td>
<td>0.006 (p=0.1449)</td>
</tr>
<tr>
<td>MFE20_50</td>
<td>0.038 (p=0.0003)</td>
<td>0.031 (p=0.0005)</td>
<td>0.019 (p=0.0095)</td>
</tr>
<tr>
<td>MFE23_25</td>
<td>0.018 (p=0.0117)</td>
<td>0.005 (p=0.1513)</td>
<td>0.004 (p=0.2183)</td>
</tr>
<tr>
<td>MFE23_50</td>
<td>0.022 (p=0.0054)</td>
<td>0.019 (p=0.0061)</td>
<td>0.009 (p=0.0781)</td>
</tr>
<tr>
<td>RBS_Score (25)</td>
<td>0.015 (p=0.0237)</td>
<td>0.007 (p=0.0925)</td>
<td>0.018 (p=0.0115)</td>
</tr>
<tr>
<td>RBS_Score (50)</td>
<td>0.015 (p=0.0226)</td>
<td>0.004 (p=0.2126)</td>
<td>0.019 (p=0.0089)</td>
</tr>
</tbody>
</table>

* See “Methods” section for the detailed definition of each computation method.
Figure 1
Figure 2

Minimum free energy of SD-rRNA interaction

- All DVU proteins
- Proteins identified under LL
- Proteins identified under FL
- Proteins identified under LS
Cellular functions

I) All DVU genes

II) LL

III) FL

IV) LS

Figure 3
Figure 5A
Figure 5B

Percent of differences (%)

Ile

Gly

Ala

Val

Thr

Pro

Leu

Ser

Arg

All  LL  FL  LS  All  LL  FL  LS  All  LL  FL  LS  All  LL  FL  LS
Figure 7

A

Stop codon

- TAA
- TAG
- TGA

Frequency (%)

All DVU proteins
Proteins identified under LL
Proteins identified under FL
Proteins identified under LS

B

Termination signal

- TAAN
- TAGN
- TGAN

Frequency (%)

All DVU proteins
Proteins identified under LL
Proteins identified under FL
Proteins identified under LS